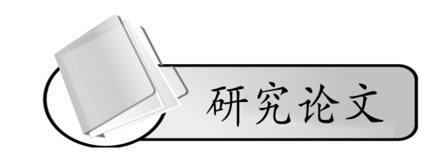


全球数据科学课程建设现状的实证分析*

朝乐门[1,2] 杨灿军[2] 王盛杰[2] 赵俊鹏[2] 许梦甜[2]
[1](数据工程与知识工程教育部重点实验室(中国人民大学) 北京 100872)
[2](中国人民大学信息资源管理学院 北京 100872)

摘要:【目的】在调查分析全球数据科学课程建设现状的基础上，提出数据科学课程的共性特点、主要挑战及解决对策。【方法】采用实证研究方法和内容分析方法调查分析国内外数据科学课程的建设现状、成功经验与存在问题。【结果】提炼出全球数据科学课程的共性以及数据科学与其他相关课程之间的差异性。【局限】对数据科学人才培养的讨论主要聚焦于课程建设层面，而对专业层面的讨论不多。【结论】本文提出数据科学课程建设中的 10 个核心问题及其解决方案。

关键词: 数据科学　大数据　课程设计
分类号: TP393

1 引　言

在大数据时代，数据科学是现代人才必备知识与技能。作为一门新兴科学，数据科学已成为高校课程设计与教学改革的主要关注点之一。在国外，哈佛大学、麻省理工学院、哥伦比亚大学、纽约大学等著名学府纷纷开设数据科学相关课程，引起全球相关专业师生的密切关注。在国内，中国人民大学、清华大学、北京大学等高等院校也先后启动数据科学专业和(或)课程建设工作。但是，就目前而言，除了少量介绍特定课程的建设经验的论文[1]之外，对国内外数据科学课程的系统调研和专题探讨仍属空白。

本文主要从课程建设层面探讨国内外数据科学课程建设中的经验与不足，分析数据科学课程教学改革中的主要问题与对策，最终探讨了数据科学课程的专业地位及在其他相关课程教学改革中的重要作用。

2 相关工作

为了收集原始数据，笔者通过互联网搜索和专家访谈的方法较为系统性地调研了国内外具有代表性的数据科学课程，如表 1 所示。在此基础上，采用内容分析法，分析数据科学课程的详细内容及元数据，以便进行深层次的讨论。

从表 1 可看出，数据科学课程的建设大约从 2011 年开始，率先在佛罗里达大学、加州大学伯克利分校、哥伦比亚大学等著名学府开设。目前，名称中含有“数据科学”或“Data Science”字样的课程群可以进一步细分为三种，如图 1 所示。

(1) 以数据科学的“理论基础”为中心的课程：主要讲解学习数据科学课程之前需要具备的知识准备，处于数据科学课程链的上游，一般很少涉及数据科学本身的知识。例如，John Paisley 于 2015 年在哥伦比亚大学开设的《面向数据科学的机器学习》(Machine Learning for Data Science)课程[2]主要讲解了机器学习知识；再如 MIT 的 Eric Grimson 教授等开设的《计算思维与数据科学导论》(Introduction to Computational Thinking and Data Science)[3]则侧重于对统计学知识的讲解。相对于其他课程，数据科学对统计学、机器学

通讯作者：朝乐门，ORCID: 0000-0002-4290-0918, E-mail: chaolemen@ruc.edu.cn。
*本文系国家社会科学基金项目“数据连续性的实现方法与保障机制研究”(项目编号：15BTQ054)的研究成果之一。

表 1　数据科学的课程调研

课程名称	年份	形式	学校	开课教师	选课要求
Data Science: Large-scale Advanced Data Analysis	2011	面授	佛罗里达大学	Daisy Zhe Wang	硕士
Data Science and Analytics Thought Leaders	2012	面授	加州大学伯克利分校	Ram Akella 等	不限
Introduction to Data Science	2012	面授	哥伦比亚大学	Rachel Schutt	不限
Introduction to Data Science	2013	面授	谢菲尔德大学	Paul Clough	数据相关/硕士
Data Science(Coursea)	2014	网授	约翰·霍普金斯大学	Roger D. Peng 等	不限
Executive Data Science(Coursea)	2014	网授	约翰·霍普金斯大学	Roger D. Peng 等	不限
Data Science at Scale (Coursea)	2014	网授	华盛顿大学	Bill Howe	不限
Data Science	2014	面授	哈佛大学	Rafael Irizarry 等	本科
Intro to Data Science	2014	面授	纽约大学	Brian D'Alessandro	不限
大数据科学与应用系列讲座(MOOC 学院)	2015	网授	清华大学	李军	不限
Foundations of Data Science	2015	面授	加州大学伯克利分校	John DeNero	不限
Data Sciences Basic	2015	面授	美国东北大学	Akira Suzuki	不限
Fundamentals of Data Science	2015	面授	慕尼黑大学	Goeran Kauermann	统计与科学相关
A Practical Approach to Data Science	2016	面/网授	哈佛大学	Ramon Mata-Toledo	不限
Introduction to Computational Thinking and Data Science (edx)	2016	网授	麻省理工学院(MIT)	Eric Grimson 等	不限
Process Mining: The Practice of Data Science (Coursea)	2016	网授	埃因霍芬理工大学	Wil van der Aalst	硕士
Data Science	2016	面授	法国圣艾蒂安大学	Marc Sebban	不限
Fundamentals of Data Science	2017	面授	牛津大学	Julian Gallop	不限
数据科学	2017	面授	中国人民大学	朝乐门	不限
Data Science	不详	面授	伦敦大学	Aysha Chaudhary	数据相关/硕士

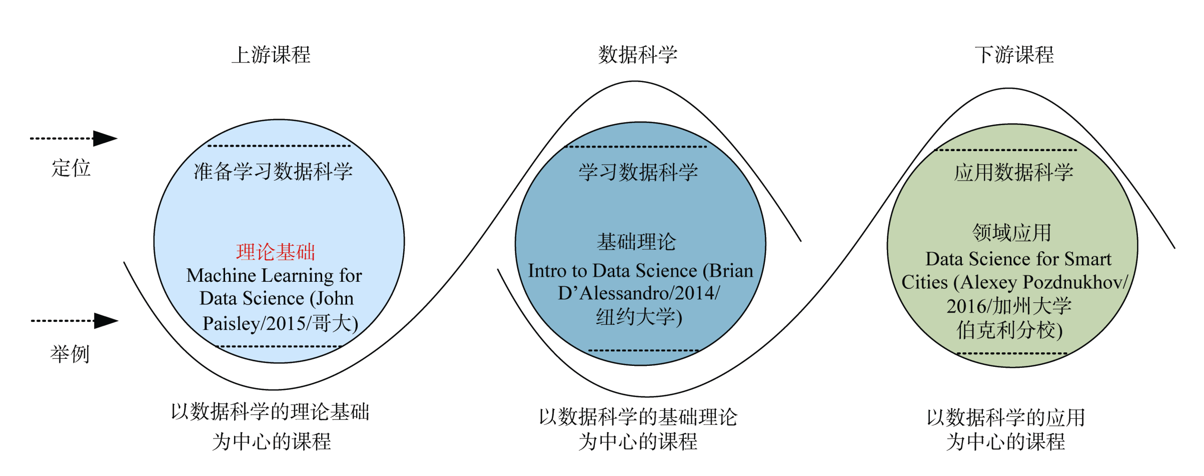


图 1　数据科学的课程链

和数据可视化等基础理论之间的耦合度较高，学习难度较大。因此，此类课程可被视为是数据科学课程的先修课程，其意义在于为学生更好掌握数据科学的知识奠定基础。

(2) 以数据科学的“基础理论”为中心的课程：主要讲解数据科学本身的理念、理论、方法、技术、工具和(或)最佳实践应用，位于数据科学课程链的核心。Brian D'Alessandro 于 2014 年在纽约大学开设的《数据科学导论》(Intro to Data Science)[4]主要讲解了数据科学的定义、方法和应用等基本问题；朝乐门在中国

人民大学开设的《数据科学》课程主要讲解数据科学的理念、理论、方法、技术、工具和典型应用，并突出了数据产品开发能力的培养。

(3) 以数据科学的“领域应用”为中心的课程：主要讲解数据科学对某一个学科领域的影响及其应用方法论或(和)最佳实践，处于数据科学课程链的下游。Alexey Pozdnukhov 于 2016 年在加州大学伯克利分校开设的课程《智慧城市中的数据科学》(Data Science for Smart Cities)[5]主要讲解如何基于数据科学解决智慧城市学科领域的数据处理工作。

值得一提的是，本文在课程建设层面重点讨论“以数据科学的基础理论为中心的课程”的建设与教学改革问题，并未涉及数据科学与大数据技术专业的建设等专业建设层面的问题。

3 共性与特色

通过相关课程的调查分析发现，数据科学的课程建设涉及两个基本问题：一是课程建设中的共性问题——虽然不同课程的建设在细节上有所不同，但它们表现出一些共性特点；二是数据科学与其他相关课程之间的差异性——数据科学课程的建设应具备的、区别于其他课程的特色属性。但是，这些共性和特色也正是数据科学课程建设与教学改革的出发点，应给予高度关注。

(1) 教学难度偏高。无论从教师的授课角度还是从学生的学习视角看，数据科学课程确实具有较高的难度系数。通过 MOOC 平台开放的数据科学课程看，虽然关注或注册人数很多，但是坚持完成全部学习和实验环节的比较少，最后能通过考试的学员人数更少。目前开设数据科学课程的学校都是哈佛、MIT、哥大等国际顶级学府，而多数一般高等院校尚未开设此课程，也在一定程度上说明该课程的教学与学习难度。从根本上讲，数据科学课程建设与改革中存在的难度系数偏高的主要原因在于：

①理论基础的跨学科性：数据科学的理论基础是统计学、机器学习和数据可视化，而这些课程的教学与学习本身就很难。

②基础理论的前沿性：数据科学的主要内容仍处于不断变化或尚未达成共识的阶段，其教与学的活动均需要不断跟踪相关理论与实践的最新动态。

③经验与材料的不足：数据科学课程的开设时间不长，可借鉴的经验、优秀教材及教学辅助材料相对匮乏。

(2) 课程定位的科学性。已开设的数据科学课程都表现出与其他传统课程的差异性。虽然数据科学中的数据统计、机器学习、数据挖掘、数据管理等内容与传统课程之间存在一定的交叉关系，但是也表现出培养目标上的区别性。例如，Ramon A. Mata-Toledo 等在哈佛大学开设的《数据科学的实用方法》(A Practical Approach to Data Science)[6]课程中包含对数据挖掘的讲解，但其教学目的并不在于系统讲解数据挖掘的全部内容，而是从数据科学视角，有选择性地讲解数据挖掘的部分知识点，突出的是面向数据科学的数据挖掘。数据科学与其理论基础课程教学之间的区别在于：

①教学目的不同：数据科学课程的教学目的并不是重新讲一遍统计学、机器学习、数据挖掘等课程的全部知识，而是为数据科学知识的学习、兴趣培育和动手能力的培养为教学目的，有选择性地讲解理论基础中的部分知识点。

②讲解视角和出发点不同：数据科学的教学不应是从统计学、机器学习角度诠释自身的内容，而是从数据科学角度讲解和组织理论基础中的知识点，做到有选择地裁剪理论基础，重视问题导向的课程设计。

③继承与创新：数据科学课程的建设不仅需要继承机器学习和统计学等理论基础类课程的建设经验，而且更需要融入一些自己的特色。例如，通过对 Anscombe 的 4 组数据(Anscombe's Quartet)[7]的介绍，讲解可视化方法、统计学和机器学习在数据科学中的互补性优势。

(3) 实战能力的培养。数据科学是一门实战性很强的课程。数据科学领域奠基之作——《Doing Data Science: Straight Talk from the Frontline》[8]、《Practical Data Science with R》[9]等畅销书的书名可以证明这一点。一些著名大学开设的课程名称也反映出数据科学的实战性，如哈佛大学《数据科学实用方法》(A Practical Approach to Data Science)、埃因霍芬理工大学《过程挖掘：数据科学实战》(Process Mining: The Practice of Data Science)、约翰·霍普金斯大学《数据科学毕业项目》(Data Science Capstone)和清华大学的《大数据科学与应用系列讲座》。从课程设计和教学实施视角看，数据科学课程的实战性主要体现在：

①加入实战应用案例分析或大作业：多数课程都引入一些特别能吸引学生兴趣的实战应用案例，如波士顿房价分析、美国总统大选预测等。中国人民大学朝乐门的《数据科学》课程特别鼓励学生通过参与开源项目和竞赛项目锻炼自己的实际动手能力。

②有实战经验的数据科学家亲自讲解或参与讲解：Dave

Holtz(Airbnb 数据科学家)等来自产业界专家在 Udacity 上开设课程《数据科学导论》(Intro to Data Science)[10]。清华大学李军开设的《大数据科学与应用系列讲座》中特邀吴军等来自实际产业和业务部门的数据科学家或相关领域的专家学者讲解专题内容。

(4) 学生专业背景的多样性。与其他课程不同的是，数据科学课程不仅受到本专业学生的重视，更是受到其他相关管理类、社会类、经济类专业学生的高度关注。目前，绝大多数数据科学课程设计也很好地反映了这一特点——课程大纲或课程通告中对学生专业要求不做任何限制。因此，生源结构的复杂性是数据科学课程建设面临的主要挑战之一。

(5) 产学研结合度高。现有数据科学课程的另一个显著特色是产学研结合程度高，而这也成为数据科学区别于其他课程的另一个特色。数据科学课程建设中的产学研结合主要表现在:

①使用产业数据。例如，在 Dave Holtz 等开设的《数据科学导论》(Intro to Data Science)[10]中采用波士顿住房数据集，包含大波士顿地区房屋所有特征的聚合数据，包括每个地区房屋的中间值。

②来自产业的师资队伍。例如，Rachel Schutt 在哥伦比亚大学开设的《Introduction to Data Science》课程中较好地处理了学生、教师和业界专家的关系——学生既是学习者，又是知识的创造者，通过课堂讨论和课下作业的方式培养了学生的自主学习能力和批判性思考能力。Rachel Schutt 曾在 Google+数据科学家团队中工作过，具备产业界的工作经历和人脉资源。该课程还邀请来自产业界的专家做专题报告，进一步增强了课程的产学研结合性。再如，Airbnb 数据科学家 Dave Holtz 等来自产业界专家在 Udacity 上开设了一门数据科学课程——《数据科学导论》(Intro to Data Science)。

③校企合作。据报道，2016 年 5 月，新加坡国立大学与微软公司合作成立新加坡国立大学数据科学研究中心，将共同推动数据科学的课程教育[11]。

4　共识与经验

在分析数据科学课程建设的共性和特色的基础上，笔者进一步深入研究了数据科学课程建设的共识与经验。相对于共性和特色，数据科学课程建设中的共识和经验更为具体，也更具有操作性。

(1) 教学目的——培养数据科学家。数据科学的教学应与传统的数据库、数据工程等相关课程的设计不同[12]。数据科学课程的主要目的是培养数据科学家，而不是数据工程师。数据科学家与数据工程师的区别如表 2 所示。可见，数据科学家不仅需要掌握一定的数据管理能力，更重要的是开发数据产品的能力。相对于数据工程师，数据科学家更需要创造性思维和批判性思考能力。

表 2　数据科学家与数据工程师的区别

对比项目	数据工程师	数据科学家
工作重点	数据的管理	基于数据的管理/决策
基本素质	工程化/标准化/ 规范化做事能力	批判性思考、问题 意识与创造力
领域差异性	领域共性较高， 领域依赖度较低	领域差异性明显， 领域依赖度较低

(2) 教学内容——数据科学的核心理论。数据科学的基本内容包括数据科学的基础理论、数据加工(Data Wrangling or Munging)、统计学、机器学习、试验设计、数据计算、数据管理、数据分析、数据科学家工具以及数据产品开发，如表 3 所示。

表 3　典型数据科学课程的对比分析

开设课程和学校	A Practical Approach to Data Science 哈佛大学	Intro to Data Science 华盛顿大学	Intro to Data Science Udacity 平台	A Crash Course in Data Science 约翰·霍普金斯大学	Introduction to Data Science 哥伦比亚大学	Fundamentals of Data Science 牛津大学
基础理论	√	√	√	√	√	√
数据加工	√	√	√	√	√	√
统计学	√	√	×	√	√	√
机器学习	×	√	×	√	√	×
数据可视化	√	√	√	×	×	×
数据管理	√	√	×	×	√	√
数据计算	√	√	√	×	√	×
数据分析	√	√	√	×	√	√
数据科学工具	√	√	√	√	√	√

其中，基础理论、数据加工、数据可视化、数据计算和数据科学工具等往往是必选内容，而统计学、机器学习和数据管理是可选内容，根据课程的培养目标和学生所具备的知识水平决定。需要注意的是，数据产品开发是数据科学的重要内容，但目前多数学校的课程中尚未突出此部分内容。

(3) 实验环节——基于 R 或 Python 的数据科学项目。从目前开设的数据科学课程看，最为常见的是 R 语言(如哈佛大学的《数据科学的实用方法》(A Practical Approach to Data Science))和 Python 语言(如 Udacity 上开设的《数据科学导论》(Intro to Data Science))。以 R 为例，基于 R 的实验环境可以分为两种：

①直接以 R 软件或 RStutio 为基础的单机实验平台。

②基于 Spark 或 Hadoop 平台进行 R 编程，即以 SparkR 和 RHadoop 为基础的集群实验平台。

但是，已开设的数据科学课程主要采用的是以 R GUI 或 RStudio 为基础的单机实验平台，而对以 SparkR 和 RHadoop 为基础的集群实验平台的研发力度不够。

(4) 教学方式——团队合作。与其他领域的科学家不同的是，数据科学家往往以团队合作为主要工作方式[13]。因此，数据科学的教学设计中往往特别强调团队学习和协同工作能力的培养。例如，哈佛大学开设的《数据科学的实用方法》(A Practical Approach to Data Science)的最后毕业项目为以团队合作方式完成 2016 美国总统大选的预测工作。值得一提的是，约翰·霍普金斯大学的数据科学课程群中专门有一门名为《数据科学团队的构建》(Building a Data Science Team)的课程，特别强调了团队合作在数据科学项目中的重要地位。

(5) 课程定位——数据科学家的多样性。在大数据时代，数据科学家可以分为两种：专业数据科学家与专业中的数据科学家。前者是数据科学专业出身，对领域知识的掌握不一定很高；后者是领域专业出身，在掌握特定领域的理论与实践之后，再学习了一定的数据科学知识，具有很强的领域意识和能力，但对数据科学本身的掌握程度不如专业数据科学家，如表 4 所示。目前，数据科学课程的设计中也充分体现了数据科学家的这一特征。例如，华盛顿大学的《数据科学导论》(Intro to Data Science)和哈佛大学的《数据科学的实用方法》(A Practical Approach to Data Science)侧重点是培养专业数据科学家；哥伦比亚大学的《数据科学导论》(Introduction to Data Science)则以培养金融、医疗、健康领域中的数据科学家为主要目的，其培养目标侧重于专业中的数据科学家。

表 4 数据科学家的差异性

对比项目	专业数据科学家	专业中的数据科学家
成长过程	起点并非领域专家，通过学习数据科学课程直接成长为数据科学家	先已成为领域专家，然后通过学习数据科学课程逐渐成为数据科学家
知识广度(数据科学)	较小(仅限于数据科学)	较大(不仅掌握领域知识，而且还掌握数据科学)
知识深度(数据科学)	较深	较高
角色定位	指导、组织、管理、监督、评价专业中的数据科学家	配合与支持专业数据科学家
相关课程(举例)	华盛顿大学的《Intro to Data Science》；哈佛大学的《A Practical Approach to Data Science》；中国人民大学的《数据科学》	哥伦比亚大学的《Introduction to Data Science》；大数据科学与应用系列讲座

5 问题与挑战

虽然数据科学课程的建设取得了一定的成功，但也存在诸多问题，主要表现在以下 5 个方面：

(1) 对主讲人的专业背景的依赖度过高。从课程中的教学设计，尤其是教学立足点和视角看，数据科学课程建设表现除了主讲人专业背景——课程内容主要由主讲人的专业背景决定，而不是从数据科学课程本身的人才培养需求出发。目前，开设数据科学课程的教师主要来自两个领域，即统计学和机器学习。来自统计学的教师容易凸显统计学知识和思维模式，强调统计学在数据科学中的主导地位，反之亦然。但是，在数据科学课程的教学中需要注意两个问题：

①统计学和机器学习并不是数据科学的基础理论，而

是其理论基础而已，严格地说应该在数据科学的范畴之外。因此，数据科学课程中适当设有统计学和机器学习的内容是可以的，但教学重点不能仅限于这些理论基础，应回归到数据科学本身的基础理论——数据科学的理念、理论、方法、技术、工具和应用；

②数据科学的教学中应平衡统计学和机器学习的关系，应把教学重点放在二者的互补优势以及在数据科学中如何综合运用统计学和机器学习的知识。

(2) 课程内容选择面广，缺乏系统性。数据科学课程的教学中存在一个重要问题是课程覆盖面太广，模块化程度过高，而对不同知识模块之间的关系讲解不够，课程内容显得碎片化，缺少系统性。

(3) 对学生专业差异性的关注不够。数据科学主要依存在不同领域之中，各领域中的数据科学存在一定差异性。例如金融大数据中强调的是数据的快速洞见，社会记忆大数据中强调的是数据的长久保存与可持续利用。但是，目前的数据科学课程中没有重视这种差异性。

(4) 对数据科学基础理论的讲解不够。需要注意的是，基础理论和理论基础是两个不同的概念。从目前的课程设计看，对统计学、机器学习等理论基础的讲解过多，而对数据科学本身的基础理论的讨论过少，没有很好地回答数据科学的基本理念、理论、方法、技术、工具、最佳实践是什么的问题。

(5) 教学方法的单一性。从课程教学方式看，主要还是以讲解为主，缺少必要的方法创新，如网授和面授的互补，翻转课堂的设计等一些新的教学方法尚未应用到数据科学的教学过程中，限制了教学效果和学习质量。

6 对策与启示

毋庸置疑，针对上述 5 个问题的解决思路大家都能想到一些，例如应避免与主讲人自己的专业背景的过度依赖等。但是，更应关注的是如何从根本上(而不是在表面上)解决数据科学课程建设中的挑战与困难。要想从根本上做好数据科学的教育，必须从深层次上解决上述局限性，需要回答以下 10 个核心问题——数据科学课程的教学设计与改革中的核心问题。

(1) 为什么要开设数据科学课程？数据科学已成为领域专家必备的知识和能力之一。如今，几乎所有的专家都在谈论大数据，但是部分“专家”并不是真正懂得大数据及其背后的科学——数据科学。在国内，数据科学的系统性研究仍属空白，人们只知道需要学习这门新兴科学，但并不知道如何学习。因此，开设数据科学课程非常有必要。

(2) 开设数据科学课程的时机是否成熟？笔者的调查发现，目前国内外有超过 80 个学校已经成功开设数据科学课程。另外，数据科学相关的图书、期刊、论文、实践、代表性人物也越来越多，已足以开设一门课程。

(3) 什么是数据科学的知识体系？数据科学的体系如图 2 所示，除了统计学、机器学习和数据可视化等理论基础之外，主要包括基础理论、数据加工、数据计算、数据管理、数据分析、数据产品研发以及在某一具体学科领域中的应用[14]。

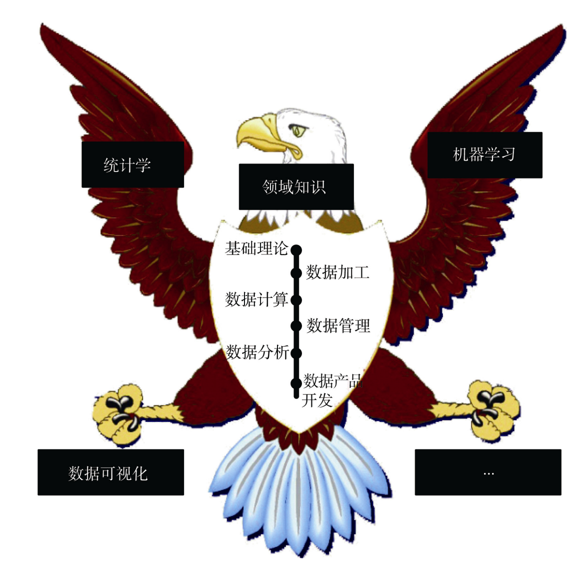


图 2　数据科学的知识体系[14]

(4) 如何设计数据科学课程？从现有课程建设经验看，数据科学课程的设计至少需要遵循四个基本原则，如表 5 所示。

表 5　数据科学课程设计的四项基本原则

序号	设计原则	应该	不应该
1	最终目标	培养数据科学家	培养数据工程师/数据管理员
2	主要特色	侧重数据科学的基础理论	侧重数据科学的理论基础
3	首要任务	培育兴趣与自学能力	讲解数据科学的全部理论
4	基本前提	统筹数据科学课程链	脱离于相关课程的独立设计

①培养数据科学家为最终目标。数据科学课程的主要培养人才是数据科学家，而不是数据工程师或数据管理员。

②侧重数据科学的基础理论为主要特色。数据科学课程的教学内容需要重视数据科学本身的基础理论，包括数据科学的研究范式、研究方法、研究技术和工具等，而不是讲解其理论基础，如统计学、机器学习、数据挖掘等。

③培育兴趣与自学能力为首要任务。数据科学家与数据科学的一个重要区别是具备高度的创造力。兴趣和信心是创造力来源。数据科学是一门快速发展的学科，不可能也没必要讲解数据科学及相关的所有理论，而应以保护和培育学生对数据学科的兴趣和信心为首要任务。实践证明，数据科学的学生多数来自非计算机和统计学类专业，而他们对统计学、机器学习本身就有距离感和恐惧感，如果引导不当，容易导致学生信心的挫败，数据科学课程的教学将以失败告终。

④统筹数据科学课程链为基本前提。数据科学课程的设计必须与相关课程统一规划，不能脱离于其他课程的设计。也就是说，数据科学课程的设计不仅需要遵循数据科学本身的特色，而且还需要注意相关课程链的设计及对本课程的影响。

(5) 数据科学课程的目标学生群体是谁？从目前的课程开设情况看，数据科学课程一般对学生专业类型和学历层次不予以限制。但是，应该注意到两个问题，如图 3 所示。

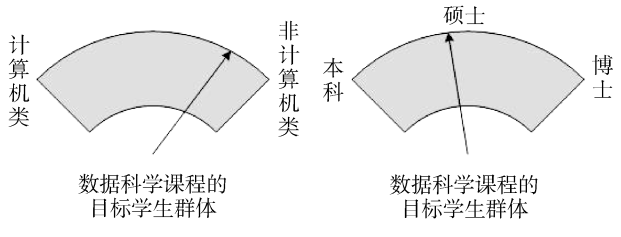


图 3　数据科学课程的目标群体

①非计算机学科领域的学生是主力军。从学科影响程度和受欢迎程度看，数据科学对传统的医疗、地理、化学、生物、管理学、金融学、社会学、图书情报、信息资源管理、历史学等非计算机领域的影响更为显著。在大数据时代，非计算机学科领域亟待思维范式的转变和动手能力的提升，学习数据科学正是他们赖以发展的主要抓手。反而，计算机学科领域由于已经学习过相关课程，往往仅需要知识系统的构建和前沿知识的导论性学习。

②硕士层次的学生是主体。国外数据科学学位项目主要集中在硕士层次，很少有本科或博士层次的学位项目，这也在一定程度上反映了数据科学课程的另一个特殊性。从数据科学本身的特点和发展现状看，数据科学课程更适合向在读研究生开设，主要原因有两个：一是数据科学对领域知识和经验的依赖度高，脱离领域问题和经验，数据科学课程教育变得枯燥而盲目，因此，数据科学课程不太适合本科生教育；另一个是数据科学本身的发展不成熟，相关理论性研究尚未健全，目前仍不适合做博士层次的教育。

(6) 数据科学课程的设计是否应该注意学生专业差异性？从目前开设课程经验看，数据科学课程的设计至少区别对待计算机及相关专业和非计算机专业，如图 4 所示。

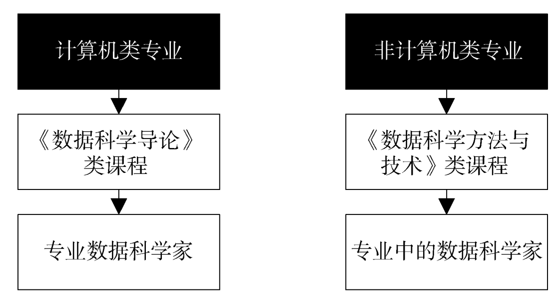


图 4　数据科学课程的专业差异性

①来自计算机及相关专业的学生需要学习的是《数据科学导论》类课程，教学设计与过程应强调“导论性”，需要给学生一个全景图，为后续进一步深入讲解做铺垫，教学目标的定位是“专业数据科学家”。

②来自非计算机专业的学生需要学习的是《数据科学方法与技术》类课程，应重视学生结合自己的领域知识和问题，采用数据科学的方法和技术进行批判性思考、动手操作和问题解决的能力，培养目的是“专业中的数据科学家”。

(7) 从事数据科学改革的瓶颈或困难在哪里？主要集中在 4 个方面，如图 5 所示。

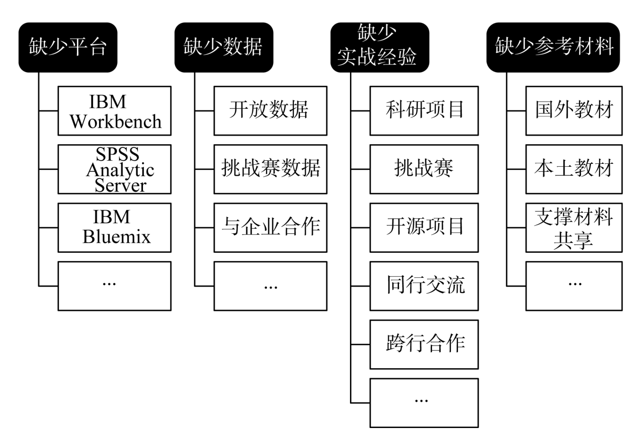


图 5　数据科学课程建设中的瓶颈

①缺少平台：数据科学课程需要数据分析，尤其是大数据管理平台。但是，很多教师和学校没有条件自己购买或搭建昂贵的实验平台。因此，数据科学的课程应积极引入 IBM Workbench、SPSS Analytic Server、IBM Bluemix 等第三方提

供的平台。

②缺少数据：数据科学课程中缺少数据是另一个瓶颈，可能的解决方案有自己抓取、与第三方合作、重用竞赛数据集、开放获取，尤其是国外政府部门开放的数据。

③缺少实战经验：应鼓励主讲人多参与科研项目、挑战赛、开源项目，重视与企业合作，加强同行交流与跨行合作。另外，建议教师们多参与相关的 GitHub 等协作平台上的开源项目以及 Kaggle 等竞赛平台上的学术竞赛。

④缺少参考材料：一直以来，数据科学课程的教学中教材选定是一个问题，主要原因是国内没有本土教材，而国外教材要么理论性太高，要么实践性太强，不适合国内教育。但是，这两年国内也有一些新的教材，如清华大学出版社出版的国内第一部系统阐述数据科学的教材《数据科学》以及《数据科学理论与实践》。

(8) 如何跟踪数据科学的最新动态？数据科学是一门不断发展和变化的学科，要求从业教师必须不断跟踪国内外相关研究进展。

①学术期刊：*Data Science Journal* (ISSN 1683-1470)、*Data Science and Engineering* (ISSN: 2364-1185)、*International Journal of Data Science and Analytics* (ISSN: 2364-415X)、*International Journal of Data Science* (ISSN: 2053-0811);

②国际会议：IEEE DSAA(IEEE International Conference on Data Science and Advanced Analytics)、ACM IKDD CODS(ACM India SIGKDD Conference on Data Sciences)、ICDSE(International Conference on Data Science and Engineering)、ICDS(The International Conference on Data Science)、Unstructured Data Science Pop-up 等;

③研究机构：伦敦帝国学院(Imperial College London)数据科学研究所、哥伦比亚大学数据科学研究所(Data Science Institute)、纽约大学的数据科学中心(NYU Center for Data Science)、加州大学伯克利分校的数据科学中心(Data Science at UC Berkeley)、全球数据科学(Data Science Global)、中国人民大学数据工程与知识工程教育部重点实验室以及一些大数据企业(如 IBM、Google、Facebook 等)的数据科学部门;

④课程资源：中国人民大学开设的数据科学 MOOC 课程；哈佛大学、麻省理工学院、斯坦福大学、纽约大学、哥伦比亚大学的数据科学及相关课程;

⑤硕士学位项目：卡内基·梅隆大学、斯坦福大学、纽约大学、加州大学伯克利分校、旧金山大学、哥伦比亚大学、佐治亚理工学院、伊利诺伊理工学院、马里兰大学和印第安纳大学等学校开设的数据科学硕士学位课程;

⑥专家学者：Alex(Sandy) Pentland(MIT 教授机器学习、人工智能与人类计算领域的知名科学家)、DJ Patil(白宫首席数据科学家)、Carlos Somohano(Data Science London 的创始人之一)、Monica Rogati(LinkedIn 高级数据科学家)、Sergey Yurgenson(哈佛教授)、Kirk Borne(2014 年被评为 IBM 大数据与分析英雄)、Hilary Mason(Fast Forward Labs 发起人，知名学者)、Yann Lecun(纽约大学数据科学中心的负责人)、Jeff Hammerbacher(曾在 Facebook 带过数据团队)、Jeremy Achin(Data Robot 创始人)、Carla Gentry(Analytical Solution 的数据科学家)、朝乐门(国内较早系统阐述数据科学专著的作者)等的个人网站(主页)、博客、Facebook 或 Twiter 等[13]。

(9) 如何处理好其他课程之间的关系？随着大数据时代的到来，原本分散在信息论、控制论和系统论等底层理论中的“数据问题”从各自学科之中独立出来，逐渐聚焦成为一门新兴学科——数据科学。数据科学对领域知识的最大影响在于将进一步抽象基础科学(如信息论、控制论和系统论等)中的“数据问题”，并使领域知识与其理论基础之间出现一门新科学，二者的研究责任与研究边界将进一步明确。也就是说，经济学、新闻学、社会学等上层理论将共享一个理论基础——数据科学，而不再直接面对信息论、控制论和系统论等底层理论，如图 6 所示。当然，数据科学课程的设计应与其他专业课程设计同步进行，需要统一的顶层设计。另外，还可以考虑采取华盛顿大学和约翰·霍普金斯大学的方式，建设一个课程群来深入讲解数据科学的内容。

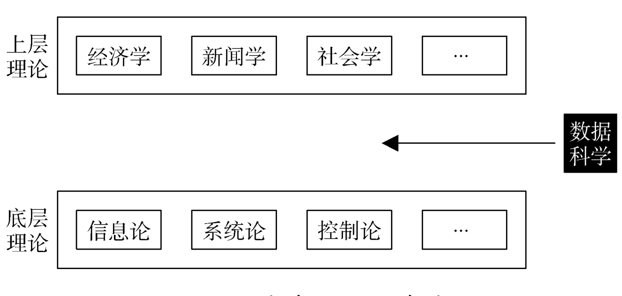


图 6 数据科学课程的学科定位

(10) 如何处理数据科学课程与数据科学专业之间的关系？数据科学课程和数据科学专业是两个不同的概念。课程不是专业的浓缩，专业也不是课程的分解。数据科学课程可以设立在数据科学专业课程体系中，也可以设立在其他专业的课程体系之中。需要注意的是，数据科学专业的培养目标往往是“专业数据科学家”，而其他专业的课程，尤其是设立在其他专业中的数据科学课程的培养目标是“专业中的数据科学家”，如图 7 所示。

总之，数据科学的课程建设不仅需要遵循课程建设的一般规律，而且还应符合数据科学本身的特殊

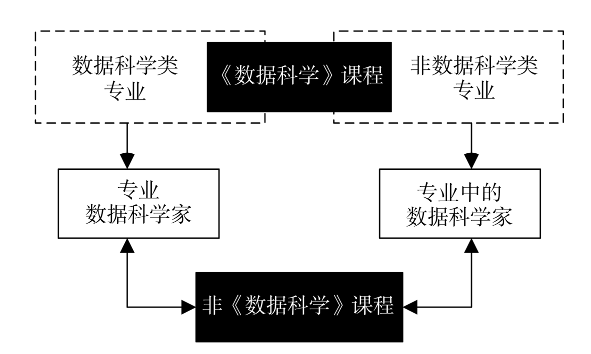


图 7　数据科学课程与数据科学专业的区别与联系

性。本文主要在教师及教学活动层面讨论了数据科学课程改革问题。但是，不能忽略另一个重要影响因素——学校及上级管理部门的正确引导和大力支持。学校及上级部门需要对数据科学课程进行顶层设计，不仅要将其纳入专业课程的必修课程目录，而且应对从事教学工作的教师给予一定的激励，如：以工作量统计时按倍数计算的方式肯定课程教学的难度、采用改进教学评价体系的方式消除教师课程改革的后顾之忧、提供教学必备条件(如平台、数据等)以解决教师课程建设中的瓶颈以及通过教学改革立项等方式提供经费支持等。数据科学课程的建设是一个系统工程，需要的是几代人的不懈努力，在此笔者也呼吁大家共同努力，为数据科学课程的教学设计与改革做出自己的贡献。

参考文献:

[1] Baumer B. A Data Science Course for Undergraduates: Thinking with Data[J]. The American Statistician, 2015, 69(4): 334-342.

[2] Paisley J. Machine Learning for Data Science [OL]. [2016-09-25]. http://www.columbia.edu/~jwp2128/Teaching/W4721/W4721Spring2015.html.

[3] Grimson E. Introduction to Computational Thinking and Data Science [OL]. [2016-08-25]. https://www.edx.org/course/introduction-computational-thinking-data-mitx-6-00-2x-4#.VL810UeUd2g.

[4] D'Alessandro B. Intro to Data Science [OL]. [2016-08-25]. http://cds.nyu.edu/course-pages/ds-ga-1001-intro-data-science/.

[5] Pozdnukhov A. Data Science for Smart Cities [OL]. [2016-08-25]. http://data8.org/smart- cities-connector/.

[6] Mata-Toledo R A. A Practical Approach to Data Science [OL]. [2016-08-25]. https://canvas.harvard.edu/courses/18032/assignments/syllabus.

[7] Anscombe F J. Graphs in Statistical Analysis [J]. The American Statistician, 1973, 27(1): 17-21.

[8] Schutt R, O'Neil C. Doing Data Science: Straight Talk from the Frontline [M]. O'Reilly Media, Inc., 2013.

[9] Zumel N, Mount J, Porzak J. Practical Data Science with R[M]. Manning Publications, 2014.

[10] Holtz D. Introduction to Data Science [OL]. [2016-08-26]. https://www.udacity.com/course/intro-to-data-science--ud359.

[11] Microsoft. NUS and Microsoft Collaborate on Data Science Education and Research [OL]. [2016-09-28]. https://news.microsoft.com/en-sg/2016/05/27/nus-and-microsoft-collaborate-on-data-science-education-and-research/#sm.0001d90kcjk04cvpsa22mvakeaiy0.

[12] Howe B, Franklin M J, Freire J, et al. Should We All be Teaching intro to Data Science Instead of Intro to Databases [C]//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. ACM, 2014: 917-918.

[13] 朝乐门. 数据科学[M]. 北京: 清华大学出版社, 2016: 41-42. (Chaolemen Borjigin. Data Science [M]. Beijing: Tsinghua University Press, 2016: 41-42.)

[14] 朝乐门. 数据科学理论与实践[M]. 北京: 清华大学出版社, 2017: 15. (Chaolemen Borjigin. Data Science: Theory and Practice [M]. Beijing: Tsinghua University Press, 2017: 15.)

作者贡献声明:

朝乐门：提出研究思路，设计研究方案，论文写作及最终版本修订；
杨灿军：数据采集、分析，部分图表的制作；
王盛杰，赵俊鹏，许梦甜：采集、清洗和分析数据。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: chaolemen@ruc.edu.cn。

[1] 朝乐门. 数据科学及相关课程调研数据表.xlsx. 数据科学及相关课程调研数据表.

[2] 朝乐门. 数据科学相关MOOC课程调研数据表.xlsx. 数据科学相关 MOOC 课程调研数据表.

收稿日期: 2017-06-12
收修改稿日期: 2017-06-15

Data Science Curriculums Around the World: An Empirical Study

Chao Lemen[1, 2]　Yang Canjun[2]　Wang Shengjie[2]　Zhao Junpeng[2]　Xu Mengtian[2]
[1](Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), Beijing 100872, China)
[2](School of Information Resource Management, Renmin University of China, Beijing 100872, China)

Abstract: [**Objective**] This paper identifies the common features of existing Data Science curriculums around the world. It also addresses the main challenges facing these courses as well as possible solutions. [**Methods**] We conducted an empirical study with the help of text analysis techniques to examine the data science curriculums from China and abroad. [**Results**] We found common features of the retrieved curriculums and the differences between them and other related courses. [**Limitations**] Our study focused on the curriculum issues, therefore, more research is needed to discuss data science as a discipline. [**Conclusions**] This paper addresses the top ten key challenges facing data science curriculum and then proposes some solutions.

Keywords: Data Science　Big Data　Curriculum Design

NISO 推出词汇表管理技术报告草案征询公众意见

美国国家信息标准组织(National Information Standards Organization, NISO)于近日就一项新的技术报告草案——词汇表管理——面向公众征询意见。该文件是 NISO 书目路线图发展项目的成果之一, 该项目始于 2013 年, 由 Andrew W. Mellon 基金会资助。书目路线图项目探讨了在全球网络环境中进行先进书目交换的可用性和采用的要求, 并在 2014 年 4 月发布的最终报告中对今后的潜在工作领域进行了优先考虑, 其中包括本技术报告中所包含的内容: 支持词汇表使用和重用的政策, 词汇表使用手册以及保存 RDF 词汇表的要求。

随着业界对分享书目信息的新环境的兴趣的增加, 有关合理的政策以及支持性基础设施的问题逐渐浮出水面。本技术报告的目的是, 为在当前过渡环境中运营词汇表的管理人员提供词汇管理的背景知识, 特别是那些不太了解相关政策和社会结构以及缺乏实践经验的运营人员。本技术报告还旨在为今后的填补工作提出一般性建议。

"讨论所有这些活动的目标是强调词汇环境中稳定性的重要性, 特别是在描述性信息移植到关联开放数据环境中时互操作的需求, " 元数据管理协会主席兼使用与重用工作组联合主席 Diane Hillmann 指出: "这些问题并不新鲜, 我们看到最近业界对这一领域的关联数据的兴趣不断增加, 希望这个技术报告能够进一步推动相关解决方案的开展。"

NISO 项目副总监 Nettie Lagace 评论说: "NISO 社区由图书馆员, 出版商, 系统和服务供应商组成, 当然这个文件也是为他们写的。但除了这些团体之外, 我们希望该文件还可以帮助许多个人和团体建立和分享书目及其他描述性数据, 以及各种组织中的知识管理人员使用词汇表解决问题。"

(编译自: http://www.niso.org/news/pr/view?item_key=9a2cb172e0cac23d1d4026fddeb2cfd28c1cbd73)

(本刊讯)